importance of appropriate imputation and how determining the mechanism of missing data informs the appropriate imputation method. A probit model using missing data dummies can effectively identify the mechanism of missing data and inform the appropriate method for imputation.

PMC16

CREATING NATIONAL WEIGHTS FOR A LARGE-SCALE, PATIENT LONGITUDINAL DATABASE

Baser O[1], Polingo L[2], Schaeffer J[3], Maguire J[4], Mummidi V[4]
[1]STATinMED Research and University of Michigan, Ann Arbor, MI, USA, [2]i3, Minneapolis, MN, USA, [3]i3, Milford, MI, USA, [4]i3, Basking Ridge, NJ, USA

Project a large-scale, patient longitudinal database to the U.S. insured population. The AHRQ Medical Expenditures Panel Survey (MEPS) was used as the basis for the adjustment methodology. MEPS are a source of data representing the cost and use of health insurance coverage, and are comprised of several large scale surveys of families, individuals, employers, and health care providers. First, we subset the data source to the study population, then used multivariate logistic regression to construct demographics and case-mix based weights that were applied to make the data similar to the national sample. The weight is derived using inverse of probability of existing in the database. To validate the weights, we randomly divided MEPS data into two parts; training set, and validating set. We used the training set to estimate the weights, then validated weights comparing standardized differences in terms of demographics and health status between the weighted and validating data sets. The following variables were used in the logistic regression: age group, gender, race, location, income levels and health status (Charlson Comorbidity Index and Chronic Conditions). i3 data were more likely to be male, older, chronic, and white ($p = 0.0000$). Adjusted weight values for the Commercial group ranged from 1 to 51 with median 1.63, Medicaid 1 to 104 with median 1.03, and Medicare 1 to 61 with median 1.07. After applying adjusted weights, standardized differences in all confounders were less than 105. National projection of a large-scale, patient longitudinal database requires adjustment from not only demographic factors but also case-mix differences related to health status. The created weights successfully balanced the population in terms of co-morbid conditions and chronic conditions as well as demographic factors.

PMC17

A METHOD FOR CONVERTING NATIONAL DRUG CODES (NDCS) TO GENERIC AND THERAPEUTIC CATEGORY CODES FOR USE IN LARGE DATABASE STUDIES OF PRESCRIPTION DRUG CLAIMS

Dickson M
University of South Carolina, College of Pharmacy, Columbia, SC, USA

OBJECTIVES: Prescription drug studies using large administrative databases usually require that NDCs be converted to generic or therapeutic categories for analysis because NDCs are structured with this information. Commercial crosswalk solutions for assigning generic and therapeutic codes to NDCs are available, but they are generally designed for applications other just converting NDCs. The cost of these comprehensive systems can be prohibitive for academic researchers conducting small to medium size projects; thus limiting research possibilities. A new system for making this conversion has been devised. **METHODS:** Because this system is designed solely for the purpose of converting NDCs to generic and therapeutic categories, it does not have many of the features found in commercially available systems (e.g., counseling notes, drug-drug interactions, patient information, etc.). However, advantages of the new system for academic researchers are its ease of use, method of delivery, simplicity, logical structure, comprehensive coverage, and free distribution. **RESULTS:** The new system provides a crosswalk from NDCs to generic entity and therapeutic category codes. Other selection criteria include dosage form, strength, and route of administration. This SQL-based system is accessible by a point-and-click user interface to select and download records of interest. Users then match the selected codes by NDC to a claims database to obtain the desired drug claims. Currently, the system includes coding for over 25,000 NDCs at the nine digit level (11 digit NDC without package size code). Additional NDCs will be added as they are identified. **CONCLUSION:** Two capabilities of this system, ability to aggregate drugs generically and therapeutically, are not found in the FDA database. In addition, the new system is retrospective whereas the FDA system is not. Academic researchers have a new tool for research using large databases of prescription claims.

PMC18

COMPARISONS OF DATA MINING ALGORITHMS FOR ADVERSE DRUG REACTIONS: AN EMPIRICAL STUDY BASED ON THE ADVERSE EVENT REPORTING SYSTEM OF THE FOOD AND DRUG ADMINISTRATION

Chen Y[1], Guo JJ[2], Patel NC[3], Steinbuch M[4], Lin XD[2], Buncher C[1]
[1]University of Cincinnati Medical Center, Cincinnati, OH, USA, [2]University of Cincinnati, Cincinnati, OH, USA, [3]University of Georgia, Augusta, GA, USA, [4]P&G Pharmaceuticals, Inc, Mason, OH, USA

OBJECTIVE: To compare the sensitivity and timing of ADRs signal early detection across the four DMAs based on the Adverse Events Reporting System (AERS) of the Food and Drug Administration. **METHODS:** The four DMAs,including the Reporting Odds Ratio (ROR), the Proportional Reporting Ratio (PRR), the Information Component (IC), and the Gamma Poisson Shrinker (GPS), are applied to retrospectively detect ten confirmed drug events combinations (DECs). The sensitivity to detect adverse events is defined as the percentage of DECs that are detected by the DMA as positive signals. The sensitivity of each DMA given different number of reports per DEC is measured as well. The timing of ADRs signal early detection is measured by comparing the index date of withdrawal (IDW) with the index date of detection (IDD). The IDW is defined as the date on which the drug was removed from the market, while the IDD is defined as a date on which the signal is significantly detected by DMAs. **RESULTS:** The estimated sensitivity to detect adverse drug event is 100% for the ROR, 90% for the PRR and the IC, and 70% for the GPS. The difference is not statistically significant. The sensitivity increases while increasing the number of reports per DEC. The average period of interval between the IDD and the IDW per DEC are approximately 10.1 quarters for the ROR, 9 quarters for the PRR, 9.9 quarters for the IC, and 4.7 quarters for the GPS, indicating the ROR is associated with the earliest signal detection among four DMAs. **CONCLUSION:** Given the overall consideration of the sensitivity and the timing of signal detection, ROR may be considered as the first choice when selecting DMAs to conduct ADRs signal detection. The findings need to be further validated in a prospective context.